We claim:

1. A method for improving information retrieval, classification, indexing, and summarization, comprising:

identifying a compound document as a coherent body of hyperlinked material on a single topic as created by a number of collaborating authors;

analyzing the content and structure of the compound document to find a preferred entry point for the compound document;

processing the compound document as a whole, including at least one of indexing, classification, and retrieval; and

processing the compound document from the entry point, including at least one of creating at least one of presentation of results from retrieval, summarization, and classification.

2. The method of claim 1 wherein the body of hyperliked material includes at least one of: the internet, an intranet, and a digital library.

3. The method of claim 1 wherein the body of hyperlinked material is distributed over a plurality of URLs.

4. The method of claim 1 wherein the identifying includes observing the results of a number of heuristics run on the body of hyperlinked material and related hyperlinks.

5. The method of claim 4 wherein the heuristic includes identifying hyperlinks that link within the same directory and include a sufficient quantity of common anchor text.

6. The method of claim 4 wherein the heuristic includes identifying hyperlinks that contain linguistic structures that indicate relationships between parts of a document including at least one of a list of page numbers, and the terms "next", "previous", "index", "contents", and their non-English equivalents.

7. The method of claim 4 wherein the heuristic includes identifying external hyperlinks to the same places.

8. The method of claim 4 wherein the heuristic includes identifying at least one of: similar creation dates and similar last-modified dates.

9. The method of claim 4 wherein the heuristic includes identifying individual URLs having similar structure indicating an order of inclusion in the compound document.

10. The method of claim 4 wherein the heuristic includes identifying a link structure of "wheel" form.

11. The method of claim 1 wherein the analyzing includes observing the results of a number of heuristics run on the component document and related hyperlinks.

12. The method of claim 11 wherein the heuristic includes identifying specific filenames that define the entry point, including at least one of: "index" and "default" .

13. The method of claim 11 wherein the heuristic includes identifying a particular component document in the compound document as the entry point because the component document has several in-links.

14. The method of claim 13 wherein the in-links are from outside the compound document.

15. The method of claim 11 wherein the heuristic includes identifying a particular component document in the compound document as the entry point because the component document has several out-links.

16. The method of claim 11 wherein the heuristic includes determining a measure of vector distances along intra-document links between a particular component document and all other component documents in the compound document.

17. The method of claim 11 wherein the heuristic includes determining whether a URL has links pointing to longer URLs having common directory components followed by different ending directory components.

18. The method of claim 17 wherein the ending directory components contain specific identifying information.

19. The method of claim 11 wherein the analyzing includes combining the results of a number of heuristics run on various component documents in the compound document, wherein the results include numerical scores and the combining includes a weighted averaging of the numerical scores into an overall score, and the maximum overall score determines the preferred entry point.

20. A system for improving information retrieval, indexing, and summarization comprising:
a compound document identifier that detects a coherent body of hyperlinked material on a single topic as created by a number of collaborating authors;
an analyzer that finds a preferred entry point for the compound document according to the content and structure of the compound document; and
a compound document processor that performs at least one of indexing, classification, and retrieval, for the compound document as a whole, and then performs at least one of creating at least one presentation of results from retrieval, summarization, and classification.

21. A computer program product instantiated on a computer-readable medium, comprising:

a first code means for identifying a compound document as a coherent body of hyperlinked material on a single topic as created by a number of collaborating authors;

a second code means for analyzing the content and structure of the compound document to find a preferred entry point for the compound document; and

a third code means for processing the compound document, wherein the processing includes at least one of creating at least one presentation of results from retrieval, summarization, and classification.